Blair Chen

github.com/chenblair | blairychen@gmail.com | 408.533.3472

EDUCATION

Carnegie Mellon University — School of Computer Science — Pittsburgh, PA

MS in Machine Learning | GPA 4.18 / 4.0 — Dec 2021

Coursework: Machine Learning for Large Datasets, Probabilistic Graphical Models, Convex Optimization, Deep Learning, Deep Learning Systems

BS in Artificial Intelligence | GPA 4.0 / 4.0 — Dec 2020

SKILLS

Python3 / Jupyter / C++ / Bash / Elixir / Javascript
pytorch / transformers / scikit-learn / numpy / pandas / plotly

EXPERIENCE

Google Document AI — Sunnyvale, CA

ML Software Engineer — Jan 2022 - Present

- Owned Expense Processor entity extraction model with $Xm of committed revenue
- Improved model F1 score by over 24% improvement, added support for 2 new languages, and expanded entity set by 20% through integration and labeling of novel data sources
- Standardized data preprocessing, storage, and labeling practices across team to address privacy concerns
- Implemented techniques to handle low data and high data regimes, including large language model pre-training, data synthesis pipelines, sub-dataset selection, and selective labeling

TruEra Intelligence Team — Redwood City, CA

Machine Learning Intern — May 2021 - Aug 2021

- Researched local and global interpretability metrics and visualizations for huggingface NLP models
- Developed a novel, linear-time approximation algorithm to quantify token interaction
- Implemented production Python code to automatically generate Jupyter notebook visualizations in a notebook experience product for data scientists

Microsoft Speech Team — Redmond, WA

Software Engineering Intern — May 2020 - Aug 2020

- Optimized hyperparameter scheduling on a *distributed* neural network training algorithm for speech recognition
- Invented a custom scheduler written in Pytorch DDP to provide reasonable baselines on unseen datasets
- Discovered a hyperparameter scheduling strategy that improved training speed by 11%
- Designed and implemented a novel hyperparameter scheduling framework for flexible adaptive scheduling (patent pending)

Airbnb Content Intelligence Team — San Francisco, CA

Software Engineering Intern — May 2019 - Aug 2019

- Built an end-to-end system for fake account detection on an intelligence pipeline, including a backend Elixir server, a React labeler frontend, and an machine learning backend that labeled 10,000s of fake accounts a day
- Trained a multilayer Bert-LSTM model to flag fake accounts automatically, automating 10 hours a week of operations work

RESEARCH

CMU Machine Learning Department — Sep 2019 - May 2021

Advisor: Dr. Louis-Philippe Morency

- Ziyin, L., Chen, B., Wang, R., Liang, P. P., Salakhutdinov, R., Morency, L. P., & Ueda, M. (2020). Learning not to learn in the presence of noisy labels. *arXiv preprint arXiv:2002.06541.*
- Chen, B., Ziyin, L., Wang, Z., & Liang, P. P. (2020). An investigation of how label smoothing affects generalization. *arXiv preprint arXiv:2010.12648.*
- Uncovered un-interpretability of attention values for overparameterized RNN training regimes

HONORS

Phi Beta Kappa — USA Computing Olympiad (Platinum) — USA Math Olympiad